Item Sum Double-List Technique:

An Enhanced Design for Asking Quantitative Sensitive Questions

Ivar Krumpal

University of Leipzig, Germany

Ben Jann

University of Bern, Switzerland

Martin Korndörfer

University of Leipzig, Germany

Stefan C. Schmukle

University of Leipzig, Germany

Word count: 8454

Date: September 14, 2017

Author Note: Martin Korndörfer is now at Salus gGmbH, Forensic ambulance, Am Kirchtor 20b, 06108 Halle/Saale, Germany

Acknowledgments: In this paper we make use of data of the LISS (Longitudinal Internet Studies for the Social sciences) panel. The LISS panel data were collected by CentERdata (Tilburg University, The Netherlands) through its MESS project funded by the Netherlands Organization for Scientific Research. The data analyzed in this paper consist of the data from our study (CentERdata 2014) merged with the May 2014 distribution of the LISS background variables (CentERdata 2012). Replication materials are available from the journal's website.

Contact information: Correspondence concerning this article should be addressed to Ivar Krumpal, Department of Sociology, University of Leipzig, Beethovenstraße 15, 04107 Leipzig, Germany. E-mail: krumpal@sozio.uni-leipzig.de

Abstract

Social desirability bias is a problem in surveys collecting data on sensitive or private topics (e.g. sexual practices, health, income, deviant behavior) as soon as the respondent's true status differs from a social norm. If confronted with sensitive questions, respondents often engage in self-protective behavior, either by giving socially desirable answers or by refusing to answer at all. Such systematic misreporting or nonresponse leads to biased estimates and poor data quality. To improve the measurement of sensitive topics in population surveys, various indirect questioning techniques have been proposed in the literature. One example, for the measurement of quantitative sensitive characteristics, is the "item sum technique" (IST). In this study we propose an enhanced design for the IST: the "item sum double-list technique" (ISDLT). Compared to the original IST, the ISDLT estimator has a higher statistical efficiency given the same sample size. We first describe our enhanced design, derive prevalence and variance estimators, and show how data collected by the ISDLT can be analyzed. We then provide evidence on the empirical viability of the ISDLT based on a large-scale experimental online survey that asked respondents about their lifetime number of sexual partners and their pornography consumption.

Keywords: social desirability; sensitive questions; response bias; item count technique; item sum technique

1 Introduction

Have you ever made false statements on tax forms in order to pay less? Thinking about the time since your 18th birthday, have you ever had sex with a person you paid, or who paid you, for sex? Such sensitive questions are not uncommon in population surveys (the first question has been used in the German General Social Survey, ALLBUS; the second question is from the US General Social Survey, GSS). Can we expect that respondents answer honestly to such questions? Cumulative empirical evidence shows that the answer is "no". Respondents often engage in self-protective behavior when asked about norm violations, either by giving socially desirable answers that do not reflect the truth (underreporting of socially undesirable behaviors and overreporting of socially desirable ones) or by refusing to answer at all (Tourangeau & Yan, 2007; Krumpal, 2013). Such systematic misreporting or nonresponse leads to biased estimates and poor data quality of a survey study.

Various data collection strategies to increase respondents' cooperation and improve the validity of self-reports in sensitive surveys have been proposed in the literature. One approach to overcome social desirability bias and elicit more honest answers is to use questioning techniques that increase the anonymity of the question-and-answer process. Such "dejeopardizing techniques" (Lee, 1993) protect the respondent's privacy by breaking the direct link between an individual answer in the survey and the "true" answer to the sensitive question. The most prominent of these techniques is the randomized response technique (RRT; Warner, 1965) that uses a randomizing device to conceal the true individual answers to the researches by introducing misclassification to the respondents' responses.

Another example is the item count technique (ICT; Droitcour et al., 1991; Biemer, Jordan, Hubbard, & Wright, 2005; Wolter & Laier, 2014), also known as the unmatched count technique (Dalton, Wimbush, & Daily, 1994; Coutts & Jann, 2011) or the list

experiment (Blair & Imai, 2012; Glynn, 2013). Compared to the RRT, the ICT has the advantage that no randomizing device has to be operated by the respondents; the ICT is thus easier to administer in an empirical survey study (another alternative to the RRT without the need for a randomizing device is the crosswise model; Yu, Tian, Tang, 2008; see Korndörfer, Krumpal, & Schmukle, for an application of this model, and Krumpal, Jann, Auspurg, & von Hermanni, 2015, for a discussion of different techniques). The ICT works as follows. The researcher randomly divides the respondents into two subsamples. One subsample receives a list of innocuous questions (short list: SL). The other subsample receives the same list of innocuous questions plus the sensitive question of interest (long list: LL). All questions are binary (i.e. each question can be answered by "yes" or "no"). Respondents in both subsamples are then asked to report the number of questions that apply to them (i.e. the total number of "yes" answers), without answering each question separately. Using such a design, it remains secret whether the sensitive question applies to a respondent (unless the respondent reports that all or none of the questions in the list apply). However, the prevalence of the sensitive behavior (i.e. the proportion of respondents to which the sensitive question applies) can be estimated as a simple mean difference of the answers in the long-list subsample and the short-list subsample (for more sophisticated estimators also see Corstange, 2009; Imai, 2011; Blair & Imai, 2012).

Several experimental studies show that the ICT is more effective than standard direct questioning in eliciting more accurate self-reports of sensitive behaviors like shoplifting (Tsuchiya, Hirai, & Ono, 2007), risky sexual practices (LaBrie & Earleywine, 2000), employee misconduct (Wimbush & Dalton, 1997) or voter turnout (Holbrook & Krosnick, 2010). However, there are also studies documenting failures of the ICT in eliciting more socially desirable answers in regards to cocaine use (Biemer & Brown, 2005) and counterproductive behaviors (Ahart & Sackett, 2004). Overviews of the empirical evidence with respect to the ICT can be found in Tourangeau and Yan (2007) and Wolter and Laier

(2014). The underlying principle of counting items is simple and comprehensible to most respondents. Furthermore, the practical implementation is straightforward in both interviewer- and self-administered data collection modes.

A main drawback of the ICT, however, is its low statistical efficiency. For a given sample size, estimates obtained from the ICT have considerably larger standard errors than estimates based on direct questioning or typically also than estimates based on the RRT. The efficiency of the ICT can be improved by a smart choice of the innocuous items (e.g., by using negatively correlated items; see Glynn, 2013). Another possibility is the use of a double-list design where, instead of only answering one list, respondents in both subsamples answer a long list and a short list (Droitcour et al., 1991; Biemer et al., 2005; Coutts, Jann, Krumpal & Näher, 2011; Kirchner, Krumpal, Trappmann, & von Hermanni, 2013). That is, based on two sets of innocuous items, the double-list design applies the ICT twice for the same sensitive question, with the roles of the two subsamples flipped, and thus provides two separate estimates of the sensitive behavior. Combining the two estimates, a more efficient estimator can be obtained. Because in the double-list design respondents from both subsamples provide an answer to a long list including the sensitive question, the effective sample size is roughly doubled compared to the single-list ICT.

Although the ICT was initially developed for the measurement of dichotomous sensitive behaviors, recently, Trappmann, Krumpal, Kirchner, and Jann (2014) proposed a generalization of the ICT for the measurement of quantitative sensitive variables: The item sum technique (IST). Analogous to the single list ICT for dichotomous variables, respondents are randomly split into two subsamples. Subjects in one subsample receive a long list (LL), containing a set of innocuous quantitative items plus one sensitive quantitative item. Subjects in the other subsample are requested to answer to a short list (SL), containing the same

innocuous quantitative items, but not the sensitive item. For example, to estimate the amount of earnings from undeclared work Trappmann et al. (2014) used the following items[1]:

- "How high are your monthly costs for your apartment or your home? Monthly costs can include rent, utilities, coop and condo fees, and mortgage." (LL and SL)
- "On average, how much do you earn per month from undeclared work?" (LL only)

The amount of earnings from undeclared work remains secret at the individual level because respondents in the LL subsample only report the sum from both items and respondents in the SL subsample do not answer to the sensitive question at all. In the IST, an estimate of the amount of earnings from undeclared work $\hat{\mu}$ can be simply calculated as the mean difference of answers between the two subsamples, that is:

$$\hat{\mu} = \bar{x}_{LL} - \bar{x}_{SL}$$

where $\bar{x}_{LL}$ is the mean in the long list subsample and $\bar{x}_{SL}$ is the mean in the short list subsample. Comparing the results to direct self-reports, Trappmann et al. (2014) found substantially higher estimates of earnings from undeclared work when using the IST. Trappmann et al. (2014) also showed how to estimate the sampling variance and derived regression models for the analysis of single list IST data.

The IST is an important new development, because many research questions are quantitative in nature, but most of the dejeopardizing techniques may only be used for dichotomous data. However, similar to the ICT, the IST has rather low statistical efficiency. A possible way to improve this low efficiency is the usage of a double-list design which has, however, until now only been used to measure dichotomous sensitive behaviors. In this paper, we therefore present the item sum double-list technique (ISDLT), a generalization of the double-list approach to quantitative sensitive questions. In section 2 we describe the ISDLT procedure and show how its data can be analyzed. Sections 3 and 4 then present the

[1] It is preferable (but not strictly necessary) that all items use the same scale (e.g. euro or count; see Trappmann et al., 2014).

implementation and results of an empirical application of the ISDLT in an online survey on pornography consumption and the lifetime number of sexual partners. Section 5 concludes.

2 The Item Sum Double-List Technique (ISDLT)

In the following, we transfer the aforementioned logic of the double-list ICT to the IST to obtain an enhanced and statistically optimized design for the measurement of quantitative sensitive behaviors, the ISDLT. Compared to the single list approach, it is obvious that the ISDLT design will lead to estimators with higher statistical power given a specific sample size, because information on the sensitive behavior is collected from respondents in both subsamples and not just in one.

The ISTDL can be implemented analogous to the ICT double list variant for dichotomous items, with the difference that all items used are quantitative: Two random subsamples are generated, whose respondents either receive long list 1 (LL1) and short list 2 (SL2) or short list 1 (SL1) and long list 2 (LL2). The two short lists contain different sets of innocuous questions, i.e. SL1 ≠ SL2. LL1 contains SL1 plus the sensitive key item. LL2 contains SL2 plus the same sensitive key item. Respondents in the first subsample receive LL1 and SL2, that is, are asked to report the sum of their answers to the questions in LL1 and the sum of their answers to the questions in SL2. Respondents in the second subsample receive SL1 and LL2. Table 1 shows a simple example with just one question per short list.

Table 1

An Example of the Item Sum Double-List Technique (ISDLT)

Subsample A	Subsample B
LL1 How much do you spend per month on housing? *How much do you earn per month from undeclared work?"*	SL1 How much do you spend per month on housing?
SL2 How much do you spend per month on food?	LL2 How much do you spend per month on food? *How much do you earn per month from undeclared work?*

Note. SL = short list, LL = long list.

While, in principle, there is no restriction on the length of the lists, it is desirable to keep them as short as possible because cognitive demand of summing up the single answers increases with the number of items in the lists and statistical efficiency tends to decline. For these reasons we suggest to use just one innocuous item per list, as in the example above (see also Trappmann et al., 2014). In this case, when answering to one of the long lists that contain the sensitive item, respondents always report a sum of the sensitive item and one of the innocuous items. Thus, the true value of the sensitive item is obscured at the individual level. When answering to one of the short lists that contain one of the innocuous items, respondents answer to this item directly. Assuming that the respondents recognize and trust this offer of privacy protection, it can be hypothesized that the ISTDL will elicit more honest self-reports of the sensitive behavior than standard direct questioning.

Several approaches can be used to obtain an estimate for the sensitive question from an ISDLT design. Given are two random subsamples A and B of size N_A and N_B,

respectively. The total sample size is $N = N_A + N_B$. There are one sensitive item S and two control items C_1 and C_2. In subsample A respondents are asked for the value of the sum of S and C_1 and for the value of C_2. In subsample B respondents are asked for the value of C_1 and the value of the sum of S and C_2. Hence, there are two response variables, Y_1 and Y_2, defined as:

$$Y_{1i} = \begin{cases} S_i + C_{1i} & \text{if } i \in A \\ C_{1i} & \text{if } i \in B \end{cases} \quad \text{and} \quad Y_{2i} = \begin{cases} C_{2i} & \text{if } i \in A \\ S_i + C_{2i} & \text{if } i \in B \end{cases}$$

From Y_1 and Y_2 we can obtain two separate estimates for the population mean of S:

$$\hat{E}_1[S] = \bar{Y}_1^A - \bar{Y}_1^B = \frac{1}{N_A}\sum_{i \in A} Y_{1i} - \frac{1}{N_B}\sum_{i \in B} Y_{1i}$$

and

$$\hat{E}_2[S] = \bar{Y}_2^B - \bar{Y}_2^A = \frac{1}{N_B}\sum_{i \in B} Y_{2i} - \frac{1}{N_A}\sum_{i \in A} Y_{2i}$$

Averaging across the two estimates, we obtain a joint estimate

$$\hat{E}[S] = \frac{\hat{E}_1[S] + \hat{E}_2[S]}{2} = \frac{(\bar{Y}_1^A - \bar{Y}_1^B) + (\bar{Y}_2^B - \bar{Y}_2^A)}{2}$$

The sampling variance of $\hat{E}[S]$ can be obtained by joint estimation of the variance matrix of the four means and then applying standard rules for linear combinations of random variables (see, e.g., Mood, Graybill, & Boes, 1974, p. 178-179).

In analogy to the above approach, regression coefficients for S with respect to a covariate vector $X_i = (X_{1i}, X_{2i}, \ldots, X_{ki})$ (including a constant) can be estimated by fitting two separate least-squares models,

$$Y_{1i} = G_i X_i'\beta + X_i'\gamma_1 + \epsilon_{1i} \quad \text{and} \quad Y_{2i} = (1 - G_i)X_i'\beta + X_i'\gamma_2 + \epsilon_{2i}$$

where G_i is an indicator for the subsample, with $G_i = 1$ for subsample A and $G_i = 0$ for subsample B, and then averaging the β estimates from the two models. To estimate the variance matrix of the averaged β coefficients, an estimate of the joint variance matrix across the two separate coefficient vectors is needed, which can be obtained by the seemingly unrelated estimation approach (see Weesie, 1999).

The above procedure averages between two separate estimates, which might not be the most efficient approach (if the subsamples are of about the same size and if C_1 and C_2 are "similar", however, averaging is a reasonable choice). A potentially more efficient approach is to estimate the two regression equations simultaneously (e.g. using Zellner's seemingly unrelated regression; Zellner, 1962), while constraining the β coefficients to be the same in both equations. Furthermore, maximum-likelihood estimation can be used. Let $S_i = X_i'\beta + \epsilon_i$ and $C_{1i} = X_i'\gamma_1 + \nu_{1i}$ and $C_{2i} = X_i'\gamma_2 + \nu_{2i}$, assuming $E(\epsilon_i) = E(\nu_{1i}) = E(\nu_{2i}) = 0$ and multivariate normality of the error terms. The log-likelihood function can be written as $\ln L = \sum_{i=1}^{N} \ln \ell_i$ with

$$\ln \ell_i = G_i \ln\left[\phi\left(Y_{1i} - X_i'\beta - X_i'\gamma_1, \sigma_{\epsilon+\nu_1}, Y_{2i} - X_i'\gamma_2, \sigma_{\nu_2}, \rho_{\epsilon+\nu_1,\nu_2}\right)\right] + (1 - G_i) \ln\left[\phi\left(Y_{2i} - X_i'\beta - X_i'\gamma_2, \sigma_{\epsilon+\nu_2}, Y_{1i} - X_i'\gamma_1, \sigma_{\nu_1}, \rho_{\epsilon+\nu_2,\nu_1}\right)\right]$$

where $\phi\left(x, \sigma_x, y, \sigma_y, \rho\right)$ is the bivariate normal density of x and y with standard deviations σ_x and σ_y and correlation ρ. Since

$$\sigma^2_{\epsilon+\nu_1+\nu_2} = \sigma^2_{\epsilon+\nu_1} + \sigma^2_{\nu_2} + 2\sigma_{\epsilon+\nu_1}\sigma_{\nu_2}\rho_{\epsilon+\nu_1,\nu_2} = \sigma^2_{\epsilon+\nu_2} + \sigma^2_{\nu_1} + 2\sigma_{\epsilon+\nu_2}\sigma_{\nu_1}\rho_{\epsilon+\nu_2,\nu_1}$$

this can be simplified (in the sense of reducing the number of unknown parameters) to

$$\ln \ell_i = G_i \ln\left[\phi\left(Y_{1i} - X_i'\beta - X_i'\gamma_1, \sigma_{\epsilon+\nu_1}, Y_{2i} - X_i'\gamma_2, \sigma_{\nu_2}, \frac{\sigma^2_{\epsilon+\nu_1+\nu_2} - \sigma^2_{\epsilon+\nu_1} - \sigma^2_{\nu_2}}{2\sigma_{\epsilon+\nu_1}\sigma_{\nu_2}}\right)\right]$$
$$+ (1 - G_i)$$
$$\times \ln\left[\phi\left(Y_{2i} - X_i'\beta - X_i'\gamma_2, \sigma_{\epsilon+\nu_2}, Y_{1i} - X_i'\gamma_1, \sigma_{\nu_1}, \frac{\sigma^2_{\epsilon+\nu_1+\nu_2} - \sigma^2_{\epsilon+\nu_2} - \sigma^2_{\nu_1}}{2\sigma_{\epsilon+\nu_2}\sigma_{\nu_1}}\right)\right]$$

Since the results from the different estimation strategies are very similar for our data, we only report the maximum-likelihood results below (results from the other approaches can be found in the online supplement).

3 The Present Study

We illustrate the ISDLT using data from an online survey in the Netherlands in which we implemented the new technique. Respondents were randomly assigned to either the ISDLT design or a standard direct questioning condition, and were asked to self-report the total number of sexual partners over their lifetime and the extent of their pornography consumption over the last 14 days.

We assume that the question about the number of past sexual partners is differentially sensitive for men and women because society imposes different normative expectations on men and women: being sexually experienced is positively connoted for men, while for women, having too many sexual partners is seen as negative. As a consequence, women tend to underreport, and men tend to overreport the number of sexual partners (Smith, 1992; Tourangeau & Smith, 1996). Past survey studies based on direct self-reports yielded substantially larger estimates of the average number of sexual partners for male respondents than for females, indicating a clear gender-specific measurement bias (see Tourangeau & Smith, 1996; Liljeros, Edling, Amaral, Stanley, & Åberg, 2001). Previous research has also shown that the measurement gap diminishes under improved anonymity conditions. That is, the average number of self-reported sexual partners decreases for men and increases for

women when data collection is more anonymous (for a comparison of self-administered data collection modes versus interviewer-administered interviews see Tourangeau & Smith, 1996). Given these results we expect that the ISDLT reduces the difference between men and women in the reported number of sexual partners, compared to standard direct questioning. This is because social norms and normative expectations are less relevant in assessment conditions that guarantee a high degree of anonymity. In particular, for men we expect the ISDLT estimates to be lower than the direct-questioning estimates ("less-is-better" assumption for socially desirable behavior) and for women we expect the reverse ("more-is-better" for socially undesirable behavior).

With respect to pornography consumption, although most adults would probably not be ashamed to admit that they have consumed pornography at least once, routine and frequent consumption is still stigmatized for both men and women. Yet, most empirical studies in this field use direct questioning to collect data on the amount and frequency of pornography consumption (e.g. Lambert, Negash, Stillman, Olmstead, & Fincham, 2012; Wetterneck, Burgess, Short, Smith, & Cervantes, 2012). We examine whether the ISDLT reduces social desirability bias and thus leads to higher estimates of pornography consumption compared to standard direct questioning ("more-is-better" assumption).

3.1 Participants

We implemented our experimental survey in the context of the LISS (Longitudinal Internet Studies for the Social sciences) panel, a probability-based internet panel with monthly self-administered online questionnaires that was in the field from 2007 to 2014. The LISS panel was funded by the Netherlands Organization for Scientific Research (NWO) and maintained by CentERdata (Institute for data collection and research located at Tilburg University, the Netherlands). The sample of the LISS panel consisted of Dutch individuals aged 16 years or older and was based on a random household sample drawn from the

population register by Statistics Netherlands (CBS). If needed, drawn households were equipped with a computer and internet connection to be able to participate in the study. Furthermore, panel members received monetary compensation for each completed interview. Detailed information about the LISS panel can be found at its website (www.lissdata.nl) and in Scherpenzeel and Das (2010).

In May 2014, our study was fielded as part of the LISS panel. The questionnaire was presented to 8033 panel members, and 6546 respondents completed the questionnaire (response rate of 81.5%). Data collection was in Dutch language. At the beginning of the interviews, respondents were randomly assigned to one of the two ISDLT groups (about 40% each; $N = 2633$ and 2580) or to the direct-questioning control group (about 20%; $N = 1333$). The ISDLT condition was oversampled because the list design is statistically less efficient than standard direct questioning and larger sample sizes are needed to achieve a comparable level of statistical power (Trappmann et al., 2014). The data analyzed in this paper consist of the data from our study (CentERdata 2014) merged with the May 2014 distribution of the LISS background variables (CentERdata 2012).

3.2 Assessment of Sensitive Questions

The design of our ISDLT implementation was as follows. Each of the long lists contained the sensitive key item and an innocuous item. Respondents were requested to indicate the sum of the two answers for each long list. In contrast, each of the corresponding short lists contained just the innocuous item, which had to be answered directly. Prior to the first questions in the ISDLT format, respondents were provided an example to exercise the usage of the new technique (see the Appendix A for the exact wording of the ISDLT long list instructions; translated from Dutch).

For the number of sexual partners the questions were as follows.

- "How many different sexual partners have you had up to now?" (sensitive question)
- "How many times did you visit a restaurant last year?" (control item 1)
- "How many cultural events (e.g. movies, concerts, theater, readings) did you go to last year?" (control item 2)

Respondents in the first ISDLT group were instructed to provide a joint answer to the sensitive question and control item 1, and answered control item 2 separately. For respondents in the second ISDLT group, the sensitive question was paired with control item 2, and control item 1 had to be answered separately. Respondents in the direct-questioning group answered all three questions separately. Likewise, for pornography consumption the questions were as follows.

- "Please think of the last 14 days. On how many of these days have you been watching pornography (e.g. via the internet, DVD, or adult movie theatre)?" (sensitive question)
- "How many days did your last holiday trip take?" (control item 1)
- "How many hours did you work last week?" (control item 2)

Apart from these experimental variations, all respondents received the same questionnaire including questions on demographics, attitudes, and norms.[2]

4 Results

4.1 Number of sexual partners

Table 2 displays the maximum-likelihood estimates of the average life-time number of sexual partners, depending on data collection mode.[3] The first column shows the estimates

[2] The data set and the corresponding documentation and codebook providing details for all variables used can be downloaded from the data archive of the LISS panel: https://www.dataarchive.lissdata.nl/study_units/view/543

[3] A few apparent outliers have been excluded from the analyses presented in Table 2. We used the following outlier rules: First, we excluded cases with obvious errors, i.e. negative or unrealistically high values in the direct questioning (DQ) items and the short lists (SL). In regards to the lifetime number of sexual partners, four cases were excluded at this step. Second, observed maxima in DQ or SL were used to define exclusion

for all respondents including males and females. These overall estimates, however, are not very meaningful because we had differential expectations for males and females. Above we argued that the question asking about the total number of past sexual partners is sensitive in different directions for men and women, that is, that men tend to over-report and women tend to under-report the number of sexual partners, and that the more anonymous ISDLT should counteract these systematic response tendencies.

Table 2

Maximum-Likelihood Estimates of the Average Lifetime Number of Sexual Partners (Standard Errors in Parentheses)

	All respondents (N = 6530)	Males only (N = 3008)	Females only (N = 3522)
Item sum estimates			
– Single-list estimate 1	3.28	3.99	2.70
	(0.50)	(0.84)	(0.59)
– Single-list estimate 2	3.22	4.06	2.50
	(0.32)	(0.52)	(0.39)
– Double-list estimate (ISDLT)	3.24	4.04	2.57
	(0.24)	(0.40)	(0.27)
Direct questioning estimate (DQ)	4.02	5.44	2.84
	(0.31)	(0.64)	(0.19)
Difference ISTDL – DQ	–0.78*	–1.40^{+}	–0.26
	(0.39)	(0.76)	(0.33)

Note. Significance test of the difference ISTDL – DQ: $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ (two-sided)

When looking at the gender-specific results, for men, as expected, the ISDLT produced a (marginally significantly; p = 0.065) lower estimate of the number of sexual partners than direct questioning (second column of Table 2). However, contrary to our expectations, female respondents reported nearly the same number of sexual partners in both

thresholds for outliers in the long lists (LL). Referring to the notation in section 2, cases in LL were excluded if values for Y_1^A and Y_2^B respectively, were larger than the sum of the maxima of the single items in DQ or SL. Another five cases exceeded the threshold values. Thus, nine cases were excluded from the analyses (see the online supplement for details).

interview conditions (the difference between ISDLT and direct questioning is also negative, but clearly not significant). One explanation for this finding may be that values and normative expectations changed over the last decades or vary between different populations. Our study population, the contemporary Dutch society, is very liberal so that there may be less pressure for women to underreport the number of their sexual partners. Former studies arguing that women underreport the number of sexual partners were conducted in populations with more conservative sex morals (e.g. Tourangeau & Smith, 1996 carried out their experimental survey in Cook County, Illinois, USA in the early 1990s). In contrast, somewhat newer studies have empirically supported the assumption that men still want to appear sexually experienced and hence exaggerate the number of their sex partners in direct self-reports (e.g. Liljeros et al., 2001, who conducted their study in Sweden in the late 1990s) and that more anonymous data collection conditions reduce such systematic overreporting. Our results are consistent with these considerations and indicate that the ISDLT is successful in reducing the gender gap in reported sexual partners (the gap substantially reduces from 5.44 – 2.84 = 2.6 using direct questioning to 4.04 – 2.57 = 1.47 using the ISDLT). However, also note that the "difference-in-differences" (i.e. the difference in effect of the questioning technique between males and females) was not statistically significant (the difference-in-differences amounts to –1.13 with a standard error of 0.83 and a *p*-value of 0.17).

At last, the results also reveal how the ISDLT successfully reduced the sampling variance compared to a single-list design. The first two rows of the table display the separate single-list estimates, which have considerably larger standard errors than the combined double-list estimates.

4.2 Pornography consumption

Table 3 displays the estimates for the days of pornography consumption over the last two weeks, depending on data collection mode.[4] For direct-questioning, we obtained an estimate of 0.82 days, i.e., on average, the respondents reported having watched pornography on a bit less than one day. Broken down by gender, we see that the overall direct-questioning estimate is mostly driven by males. Females, if asked directly, reported almost no pornography consumption. The ISDLT estimate, surprisingly, was significantly lower and essentially zero for both males and females. Such a finding is unexpected, because if the ISDLT provides anonymity to the respondents such that they are more willing to provide honest answers, we would expect the ISDLT estimate for pornography consumption to be higher than the corresponding direct-questioning estimate, not lower.

[4] Again, some outliers have been excluded from the analyses presented in Table 3. We used the same outlier rules as for the question on sexual partners. First, we excluded cases with obvious errors, i.e. negative or unrealistically high values in the direct questioning (DQ) items and the short lists (SL). In regards to pornography consumption, no obvious errors were found at this step. Second, observed maxima in DQ or SL were used to define exclusion thresholds for outliers in the long lists (LL). Referring to the notation in section 2, cases in LL were excluded if values for Y_1^A and Y_2^B respectively, were larger than the sum of the maxima of the single items in DQ or SL. Three cases exceeded the threshold values and were excluded from the analyses (see the online supplement for details).

Table 3

Maximum-Likelihood Estimates of the Average Number of Days Watching Pornography over the Last Two Weeks (Standard Errors in Parentheses)

	All respondents (N = 6533)	Males only (N = 3010)	Females only (N = 3523)
Item sum estimates			
– Single-list estimate 1	–0.13	–0.06	–0.15
	(0.55)	(0.96)	(0.59)
– Single-list estimate 2	0.03	0.44	–0.41
	(0.51)	(0.83)	(0.60)
– Double-list estimate (ISDLT)	–0.04	0.22	–0.28
	(0.38)	(0.64)	(0.43)
Direct questioning estimate (DQ)	0.82	1.64	0.14
	(0.06)	(0.12)	(0.03)
Difference ISTDL – DQ	–0.86*	–1.42*	–0.41
	(0.38)	(0.66)	(0.43)

Note. Significance test of the difference ISTDL – DQ: $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ (two-sided)

4.3 Properties of the control items and their relevance for the ISDLT estimates

Our explanation for the failure of the ISDLT to give meaningful results in the case of pornography consumption is that the control items for the sensitive question on pornography consumption were not very well chosen in our study. Table 4 displays the means and variances of the sensitive questions and the control items in the direct questioning group.

Table 4

Means and Standard Deviations of the Sensitive Questions and Control Items in the Direct Questioning Group

	Mean	Standard deviation
Sexual partners:		
"How many different sexual partners have you had up to now?"	4.0	11.4
"How many times did you visit a restaurant last year?"	10.8	14.7
"How many cultural events did you go to last year?"	5.3	8.9
Pornography consumption:		
"Please think of the last 14 days. On how many of these days have you been watching pornography?"	0.8	2.3
"How many days did your last holiday trip take?"	11.0	23.4
"How many hours did you work last week?"	18.4	18.7

As can be seen, both control items for pornography consumption have much higher means and variances than the sensitive question on pornography consumption. The difference in means is not per se a problem, however, the very large difference in variances causes the ISDLT estimate to become very inefficient. In general, the lower the variance of the control items, the lower the noise introduced by the control items and the more precise the estimate for the sensitive question (in the extreme case, when the control item variance is zero, the ISDLT estimate is equivalent to a direct-questioning estimate). A low control item variance, however, also means that there is only little privacy protection. That is, the variance of the control items determines the balance between privacy protection and statistical efficiency (the covariance between the control items and the sensitive question also plays a role: a negative correlation increases efficiency, a positive correlation reduces efficiency; in our data, these correlations are close to zero). In case of our ISDLT implementation for the pornography question we are very much on the side of privacy protection: there is about a tenfold difference in standard deviations between the control items and the sensitive question. Given

such a design, no precise estimate of pornography consumption can be obtained. This can also be seen in Table 3, where the standard error of the ISDLT estimate (0.38) is much larger than the standard error of the DQ estimate (0.06).

For the question on the number of sexual partners, the variance ratio between control items and the sensitive question is much more favorable: the standard deviations of the variables are in a similar range. Hence, the ISDLT estimate for the number of sexual partners is much more efficient. Indeed, as can be seen in Table 2, the standard error of the ISDLT estimate is even lower than the standard error of the direct questioning estimate in this case (0.24 vs. 0.31; but recall that a larger sample size has been used for the ISDLT then for direct questioning).

To summarize, the high control-item variance is a clear deficiency of our implementation of the ISDLT for pornography consumption. In general, we would suggest using control items with a variance that is in a similar range as the variance of the sensitive question, as was the case for our ISDLT implementation for the question on sexual partners

However, low precision alone cannot explain why for pornography consumption the ISDLT estimate was, in fact, significantly lower than the direct questioning estimate. Since the value for pornography consumption cannot be lower than zero, this means that there is a design effect in the sense that the control items were answered differently depending on whether they were paired with the sensitive question or not. Former empirical studies evaluating the list experiment also reported perplexing results. Measuring sensitive dichotomous behavior, Droitcour et al. (1991) found that the ICT produced smaller prevalence estimates of illicit drug use than DQ. Furthermore, Biemer and Brown (2005) found ICT estimates of cocaine use prevalence to be smaller than estimates based on DQ. Some estimates were even negative in the ICT condition. They also compared ICT and DQ answers of the same respondents and found that a considerable number of them answered “none” (0) to the set of items in the ICT format but answered “yes” (1) to all items when the

questions were presented individually in the DQ format. This could be an indication of noncompliance with the list format or possible design effects (e.g. Blair & Imai 2012).

One argument found in the literature is that there is a so-called undercounting effect. Tsuchiya and Hirai (2010) discuss the respondents' tendency to indicate a smaller number of applicable items in the list format compared to the same items answered directly: "However, the number of applicable items indicated via the item count question tends to be smaller than when it is calculated from the direct 'applies/does not apply' responses to each item." Such an effect could potentially distort estimates of the sensitive behavior. However, the effect found by Tsuchiya & Hirai (2010) applies to lists of several non-sensitive dichotomous items and it is unclear whether the ISDLT with just one quantitative non-sensitive item suffers of a similar problem.

We assume that another problem might be responsible for our unexpected result for pornography consumption using the ISDLT. Suppose a respondent has a relatively high value on the control item. The respondent might then be tempted to underreport in the long-list format, because the respondent might fear that a high value will be interpreted as an indication of excessive pornography consumption. Moreover, for respondents who want to avoid any association with pornography, a rational strategy is to answer "zero" in the long list format irrespective of the true value of the control item. Since we do not know the true control item values for respondents who answered the long list, we cannot provide direct evidence for such behavior. However, Table 5 contains an analysis of the proportion of zero-answers in the different long lists and short lists. In most cases, the number of zero-answers is smaller in the long list than in the short list, as one would expect, since in the long list the values of two variables are added together. For one of the control items paired with the question on pornography consumption, however, the number of zero-answers is significantly larger in the long list than in the short list. This is a clear indication that the respondents in the long list engaged in underreporting. Similar underreporting effects could also exist for the

other items, but they may just not be strong enough to reverse the effect. Furthermore, underreporting may not necessarily only occur in form of zero-answers; it may also be that respondents with high control item values edit their answers to be more in line with a presumed "average" value.

Table 5

Proportion of Zero-Answers by Control Item (Standard Errors in Parentheses)

	Long list	Short list	Difference
Sexual partners:			
"How many times did you visit a restaurant last year?"	5.05% (0.43)	8.94% (0.56)	–3.88%*** (0.71)
"How many cultural events did you go to last year?"	13.17% (0.67)	22.82% (0.82)	–9.65%*** (1.06)
Pornography consumption:			
"How many days did your last holiday trip take?"	7.14% (0.50)	5.48% (0.45)	1.67%* (0.67)
"How many hours did you work last week?"	33.93% (0.94)	37.32% (0.94)	–3.39%* (1.33)

Note. Significance test of the difference long list – short list: $^{+}p < 0.01$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ (two-sided)

Although we cannot say much about the exact nature of the underreporting effect based on our data, it is likely that such an effect will affect the ISDLT estimate more strongly in a situation where the control item has a larger mean and variance than the sensitive question. For example, if the mean of the control item is much larger than the mean of the sensitive question, just a hand full of respondents who answer zero instead of providing the true value of the control item in the long list can introduce substantial bias to the ISDLT estimate. This is because, relative to the mean of the sensitive question, the effect of these zero-answers on the overall mean will be very large. The situation is similar if one assumes that underreporting, in general, occurs in relation to the scale of the control item. Hence, because for pornography consumption the means and variances of the control items are so much larger than the mean and variance of the sensitive question, our ISDLT estimate for

pornography consumption is very sensitive to underreporting bias. In any case, it seems advisable to use control items that have a similar scale as the sensitive question, as is the case in our ISDLT implementation for lifetime sexual partners.

4.4 Regression estimates

As indicated in the methods section above, it is possible to fit regression models to data collected by the ISDLT. For case of exposition, Table 6 displays the results from some exploratory models to explain the number of sexual partners and pornography consumption, both for our direct-questioning sample and for the ISDLT. As covariates we use gender, age, whether the respondent was in a relationship at the time of the survey, the respondent's educational level (in six levels from primary school to university; for sake of simplicity we modelled a linear effect across the levels), respondent's attitude towards pornography (i.e., whether the respondent agreed or strongly agreed with the statement "It is wrong to watch pornography"), and the perceived social norm with respect to pornography consumption (i.e., the respondent's answer to the question: "What is your estimate of the percentage of people in your entire circle of acquaintances who watch pornography?").

Table 6

Regression Results (Standard Errors in Parentheses)

	Sexual partners		Pornography consumption	
	DQ (N = 1326)	ISDLT (N = 5176)	DQ (N = 1327)	ISDLT (N = 5178)
Gender (0 = male; 1 = female)	–1.654*	–0.421	–1.101***	0.014
	(0.656)	(0.501)	(0.117)	(0.767)
Age (in years)	0.007	0.023	–0.004	–0.012
	(0.020)	(0.015)	(0.004)	(0.023)
Respondent in a relationship (0 = no; 1 = yes)	–0.054	0.108	–0.566***	0.260
	(0.726)	(0.551)	(0.130)	(0.846)
Education (from 1 = primary school to 6 = university)	0.283	0.670***	–0.090*	0.215
	(0.204)	(0.161)	(0.037)	(0.246)
Agrees or strongly agrees with statement "It is wrong to watch pornography" (0/1)	–2.165*	–2.472***	0.038	–0.288
	(0.888)	(0.664)	(0.159)	(1.013)
Estimated percentage of acquaintances who watch pornography (0–100)	0.042**	0.046***	0.029***	0.044**
	(0.014)	(0.011)	(0.002)	(0.016)
Constant	2.917^{+}	-0.853	1.711***	–1.389
	(1.622)	(1.217)	(0.290)	(1.868)

Note. DQ = direct questioning; ISDLT = item sum double-list technique. DQ: OLS regression; ISDLT: maximum-likelihood estimation (see text for more information).

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ (two-sided)

In the direct-questioning sample we find the well-known gender gap in reported sexual partners; in the ISDLT the gap is much smaller and no longer significant (the difference of the gender effect between the direct-questioning model and the ISDLT model is marginally significant with a p-value of 0.07). Furthermore, in the ISDLT sample, but not in the direct-questioning sample, education is positively related to the number of sexual partners. Finally, in both the direct-questioning sample and the ISDLT sample, a negative attitude toward pornography is negatively related to the number of sexual partners, and the perceived social norm with respect to pornography consumption is positively related to the number of sexual partners. These effects, of course, may not be causal, and causality may also flow in reverse direction (e.g., if having few sexual partners leads to a more negative attitude towards pornography).

For pornography consumption, in the direct-questioning sample we find significant effects of gender (women consuming less pornography), education (the better educated watching less), and the perceived descriptive norm (respondents who reported that a large portion of their acquaintances watch pornography also reported higher values for their own pornography consumption). Furthermore, being in a relationship seems to reduce pornography consumption. Due to the above described problem of large control item variances, the regression results for pornography consumption in the ISDLT sample are very imprecise and only a positive effect of the descriptive norm can be found.

5 Discussion

We proposed an optimized design for the measurement of quantitative sensitive characteristics, the "item sum double-list technique" (ISDLT), which is a generalization of the "item sum technique" (IST) recently proposed by Trappmann et al. (2014). The ISDLT has the advantage over the IST that it should lead to more precise estimates, because both experimental groups provide information on the sensitive item. Thus, ISDLT requires a smaller sample size than the single-list design to achieve a given level of statistical power. In the methods section of our article, we described the technique and derived suitable estimators for the analysis of ISDLT data.

In the empirical part, we presented a first test of the practical viability of the ISDLT and compared its results to conventional direct questioning (DQ). In an experimental online survey in the Netherlands, we asked sensitive questions about the respondents' lifetime number of sexual partners and pornography consumption behavior. As expected, we consistently observed smaller standard errors for the ISDLT compared to the IST estimates. We were thus able to confirm that the ISDLT indeed leads to more precise estimates given the same sample size.

However, although the results we obtained from the ISDLT for the question of the number of past sexual partners had face validity and appeared more or less consistent with the literature using alternative techniques for increasing anonymity (Tourangeau & Smith, 1996), the results for the question on pornography consumption were unexpected. As the ISDLT provides anonymity to the respondents such that they are more willing to provide honest answers, we expected the ISDLT estimate for pornography consumption to be higher than the corresponding direct-questioning estimate, not to be lower, as we observed in our data. We identified our choice of control items as the main reason for the failure of the ISDLT to reduce social-desirability bias in measures of pornography consumption. In particular, the control items for the pornography question had much larger means and variances than the sensitive question, making the ISDLT estimate imprecise and susceptible to underreporting bias, whereas for the question on the lifetime number of sexual partners the control items had similar means and variances, producing more sensible results.

For following studies using the ISDLT we thus strongly advice using control items whose means and variances are of similar magnitude as the mean and variance of the sensitive question. This way, the procedure provides credible privacy protection, but estimates do not become too unstable. Depending on situation, it may be helpful to conduct a pilot study to evaluate different sets of control questions.

Appendix A: ISDLT Long List Instructions (Translated from Dutch)

You will now receive a block with 2 questions. Each question within the block must be answered with a number. It is also possible that you answer one or both questions with '0'.

Please memorize the answer to each question or write it down on a sheet. Afterwards, please add up the numbers resulting from both answers and indicate the total result. Since we do not know your answer to each question we do not know the composition of your results.

Let us give you a brief example. Assume the following 2 questions being asked in a block.

Question 1: How many pairs of shoes do you own?

Question 2: How many pets have you had in your life?

Suppose that you have 7 pairs of shoes. In this case, you would have to memorize or write down the number 7 for the question 1.

Suppose that you have had only 1 pet in your life so far. Then you would have to memorize or write down the number 1 for that question.

Now add up the two numbers memorized or written down and indicate the total: In this case, the number 8.

In the following questions please follow the same pattern. Memorize or write down the respective answers and only indicate the result at the end of each question block.

References

Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods*, *7*, 101–114.

Biemer, P., & Brown, G. (2005). Model-based estimation of drug use prevalence using item count data. *Journal of Official Statistics*, *21*, 287–308.

Biemer, P., Jordan, B. K., Hubbard, M. L., & Wright, D. (2005). A test of the item count methodology for estimating cocaine use prevalence. In J. Kennet & J. Gfroerer (Eds.), *Evaluating and improving methods used in the National Survey on Drug Use and Health* (pp. 149-174). Rockville: Substance Abuse and Mental Health Service Administration, Office of Applied Studies.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, *20*, 47–77.

CentERdata (2012). Background Variables. *LISS Panel Project Number 1.* Available from https://www.dataarchive.lissdata.nl/study_units/view/322.

CentERdata (2014). Measuring quantitative sensitive behaviors. *LISS Panel Project Number 129.* Available from https://www.dataarchive.lissdata.nl/study_units/view/543.

Corstange, D. (2009). Sensitive questions, truthful answers? Modelling the list experiment with LISTIT. *Political Analysis, 17*, 45–63.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys. Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research, 40*, 169-193.

Coutts, E., Jann, B., Krumpal, I., & Näher, A. F. (2011). Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions. *Journal of Economics and Statistics, 231*, 749-760.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology, 47*, 817–828.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsely, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 185-210). New York: Wiley.

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly, 77*, 159–172.

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, 74, 37-67.

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association, 106*, 407-416.

Kirchner, A., Krumpal, I., Trappmann, M., & von Hermanni, H. (2013). Messung und Erklärung von Schwarzarbeit in Deutschland - Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit. *Zeitschrift für Soziologie, 42*, 291-314.

Korndörfer, M., Krumpal, I. & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology*, *45*, 18-32.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*, 2025-2047.

Krumpal, I., Jann, B., Auspurg, K., & von Hermanni, H. (2015). Asking sensitive questions: A critical account of the randomized response technique and related methods. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel & P. Sturgis (Eds.), *Improving survey*

methods: Lessons from recent research (pp. 122-136). New York: Routledge/Taylor & Francis.

LaBrie, J. W., & Earleywine, M. (2000). Sexual risk behaviors and alcohol: higher base rates revealed using the unmatched-count technique. *Journal of Sex Research, 37*, 321-326.

Lambert, N. M., Negash, S., Stillman, T. F., Olmstead, S. B., & Fincham, F. D. (2012). A love that doesn't last: Pornography consumption and weakened commitment to one's romantic partner. *Journal of Social and Clinical Psychology, 31*, 410-438.

Lee, R. M. (1993). *Doing research on sensitive topics*. London: Sage.

Liljeros, F., Edling, C.R., Amaral, L.A.N., Stanley, H. E., & Åberg, Y. (2001). The web of human sexual contacts. *Nature, 411*, 907-908.

Mood, A. M., Graybill, F. A., & Boes, D.C. (1974). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Scherpenzeel, A.C., & Das, M. (2010). "True" longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (pp. 77-104). Boca Raton: Taylor & Francis.

Smith, T. W. (1992). Discrepancies between men and women in reporting number of sexual partners - a summary from 4 countries. *Social Biology, 39*, 203–211.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions — the impact of data collection mode, question format and question context. *Public Opinion Quarterly, 60*, 275–304.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883.

Trappmann, M., Krumpal, I., Kirchner, A., & Jann, B. (2014). Item sum - A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology, 2*, 58-77.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007). A study of the properties of the item count technique. *Public Opinion Quarterly, 71*, 253-272.

Tsuchiya, T., & Hirai, Y. (2010). Elaborate item count questioning: Why do people underreport in item count responses? *Survey Research Methods, 4*, 139-149.

Warner, S. L. (1965). Randomized-response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63-69.

Weesie, J. (1999). sg121: Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Technical Bulletin, 52*, 34-47.

Wetterneck, C. T., Burgess, A. J., Short, M. B., Smith, A. H., & Cervantes, M. E. (2012). The role of sexual compulsivity, impulsivity, and experiential avoidance in internet pornography use. *Psychological Record, 62*, 3-18.

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology, 82*, 756-763.

Wolter, F., & Laier, B. (2014). The effectiveness of the item count technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency. *Survey Research Methods, 8*, 153-168.

Yu, J. W., Tian, G. L., & Tang, M. L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*, 251–263.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*, 348-368.